Transferable Adversarial Examples: Insights, Attacks & Defenses

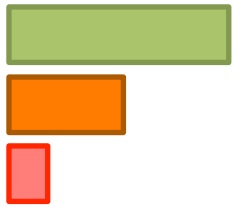
June 12th 2017

Florian Tramèr

Joint work with Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh & Patrick McDaniel

Adversarial Examples Threat Model: White-Box Attacks

ML Model



bird
tree
plane

ground truth

Loss

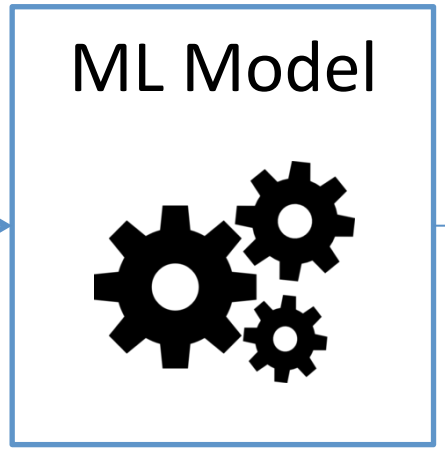
Take gradient of the loss

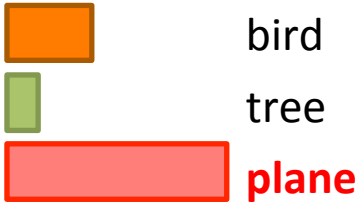
"Fast Gradient Sign Method" (FGSM)

$$r = \epsilon \cdot \mathtt{sign}\left(\nabla_x J(x, y, \theta)\right)$$

Adversarial Examples Threat Model: White-Box Attacks

+



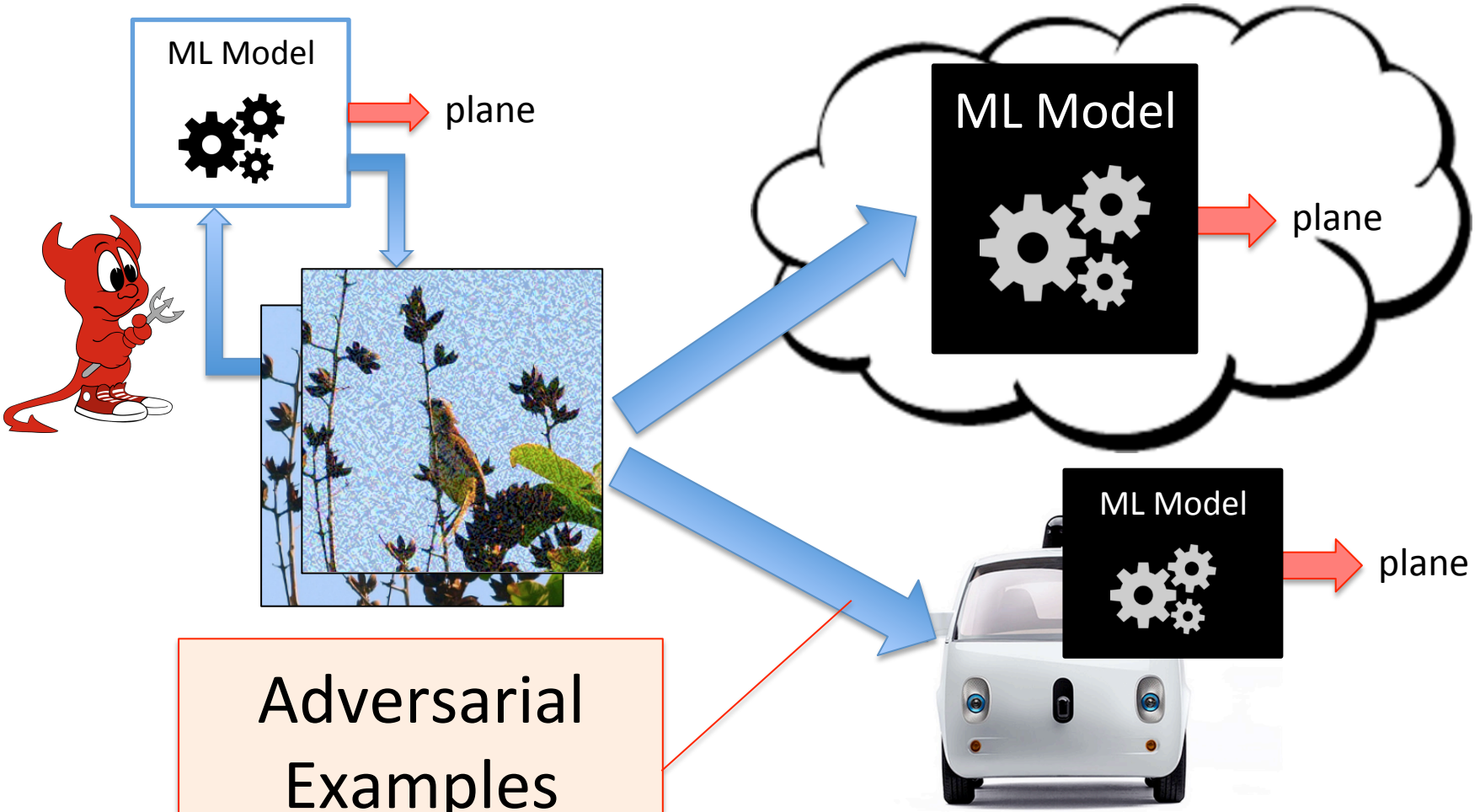


$$\|r\|_\infty = \epsilon$$

"Fast Gradient Sign Method" (FGSM)

$$r = \epsilon \cdot \mathtt{sign}\left(\nabla_x J(x, y, \theta)\right)$$

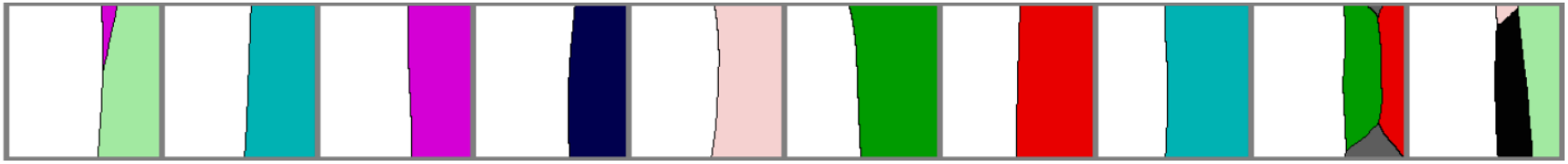
Adversarial Examples Threat Model: Black-Box Attacks



The Space of Transferable Adversarial Examples

How large is the “space” of adversarial examples?

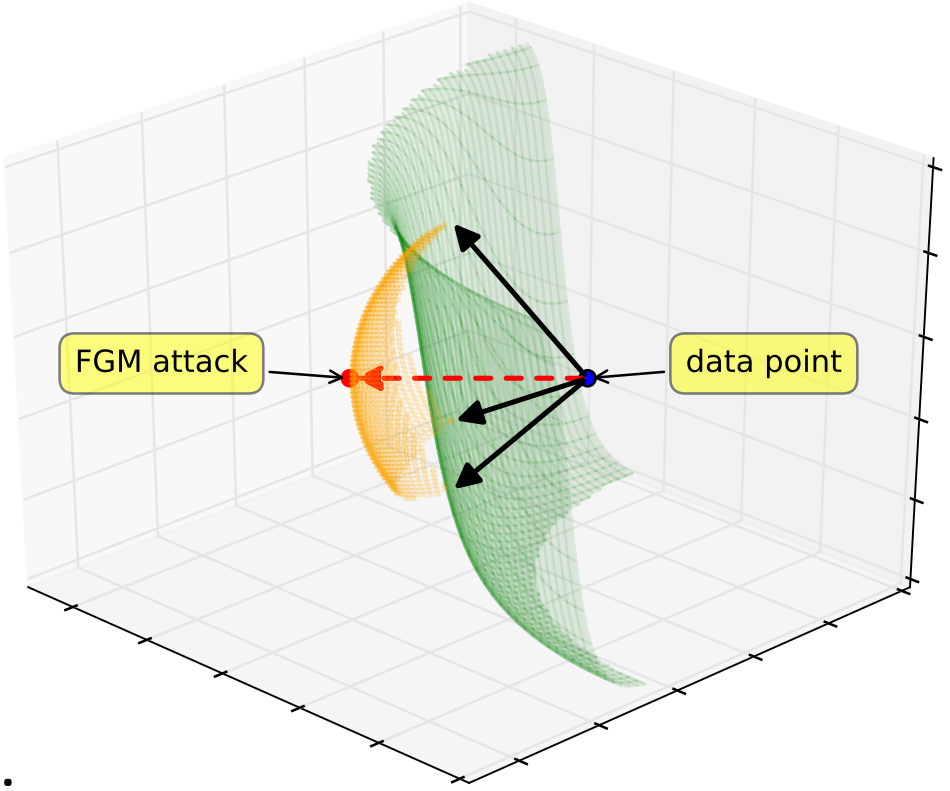
- At least 2D
  - Warde-Farley & Goodfellow 2016
  - Liu et al. 2017



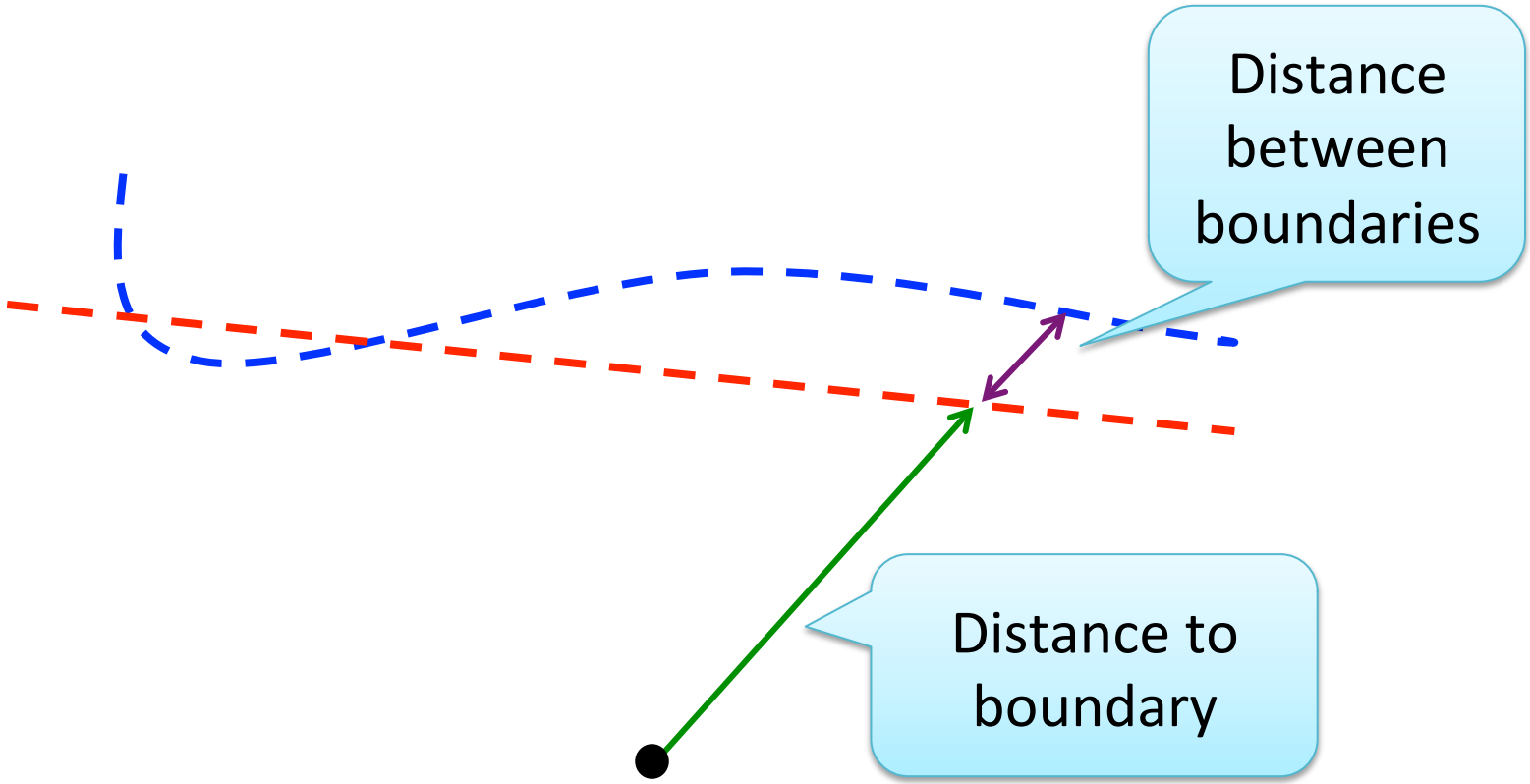
Church window plots. Warde-Farley & Goodfellow 2016

Gradient-Aligned Subspaces

- Adversarial examples form a **contiguous** subspace of **"high" dimensionality**
  - 15-45 dimensions for DNNs and CNNs on MNIST
  - *Intersection* of adversarial subspaces is also multidimensional



Decision Boundary Similarity



Decision Boundary Similarity

- Experiments with MNIST and DREBIN (malware)
  - DNN, Logistic Regression, SVM
  - 3 directions:
    - Aligned with gradient (adversarial example)
    - In direction of data point of different class
    - In random direction
- Results: In **any** direction,

  distance to boundary $\gg$ distance btw boundaries

Models are similar "everywhere"

Open Questions

- **Why** this similarity?
  - Data dependent results?
  - E.g., for a binary MNIST task (3s vs 7s) we prove:

If F_1 (linear model) and F_2 (quadratic model) have high accuracy, then there are adversarial examples that transfer between the two models

  - These adversarial examples also transfer to DNNs and CNNs but we can't prove this is inherent…

Transferability and Adversarial Training

Adversarial Training

ML Model

bird

Loss

take gradient
(FGSM)

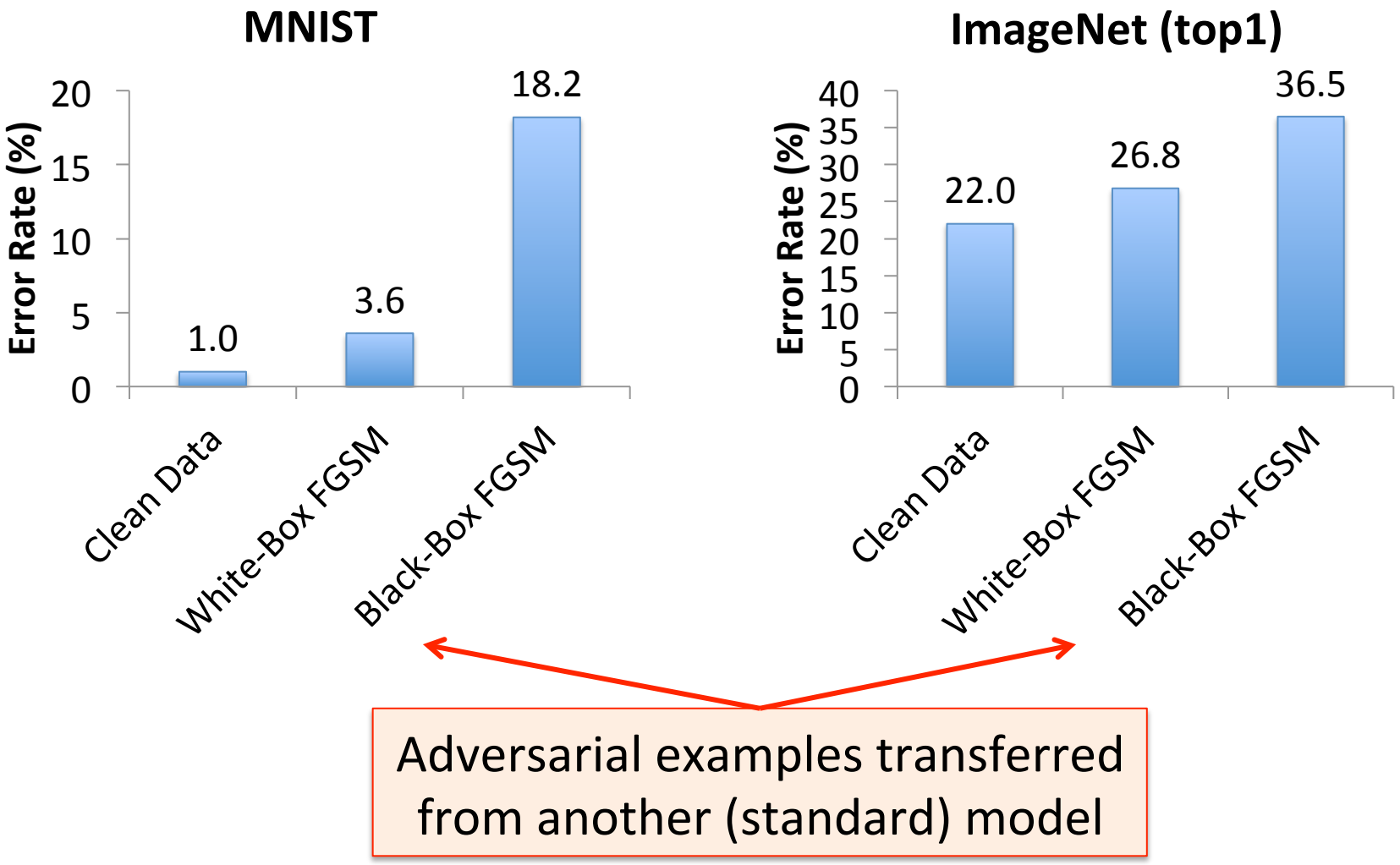
ML Model

plane

Loss

Repeat

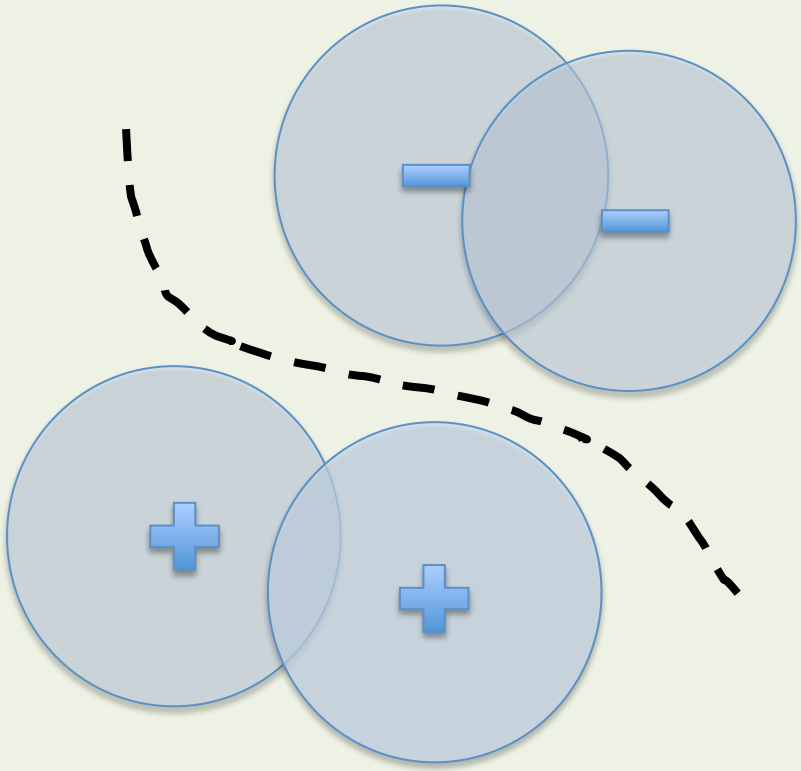
Attacks on Adversarial Training



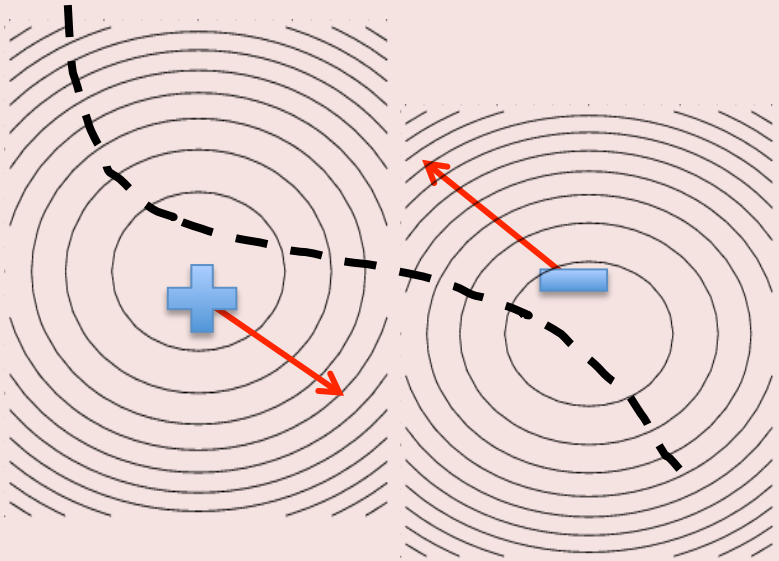
Gradient Masking

- How to get robustness to FGSM-style attacks?

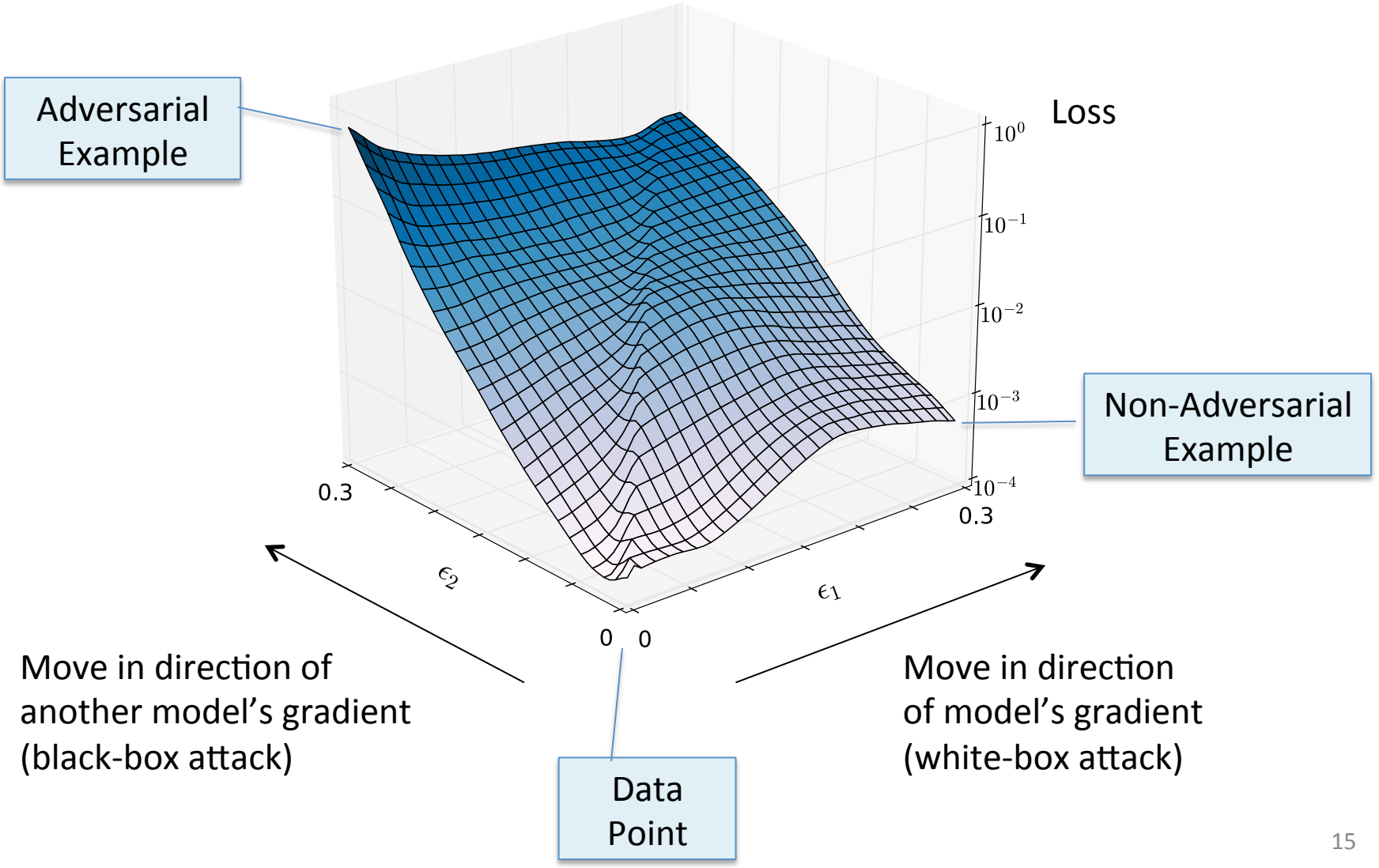
Large Margin Classifier



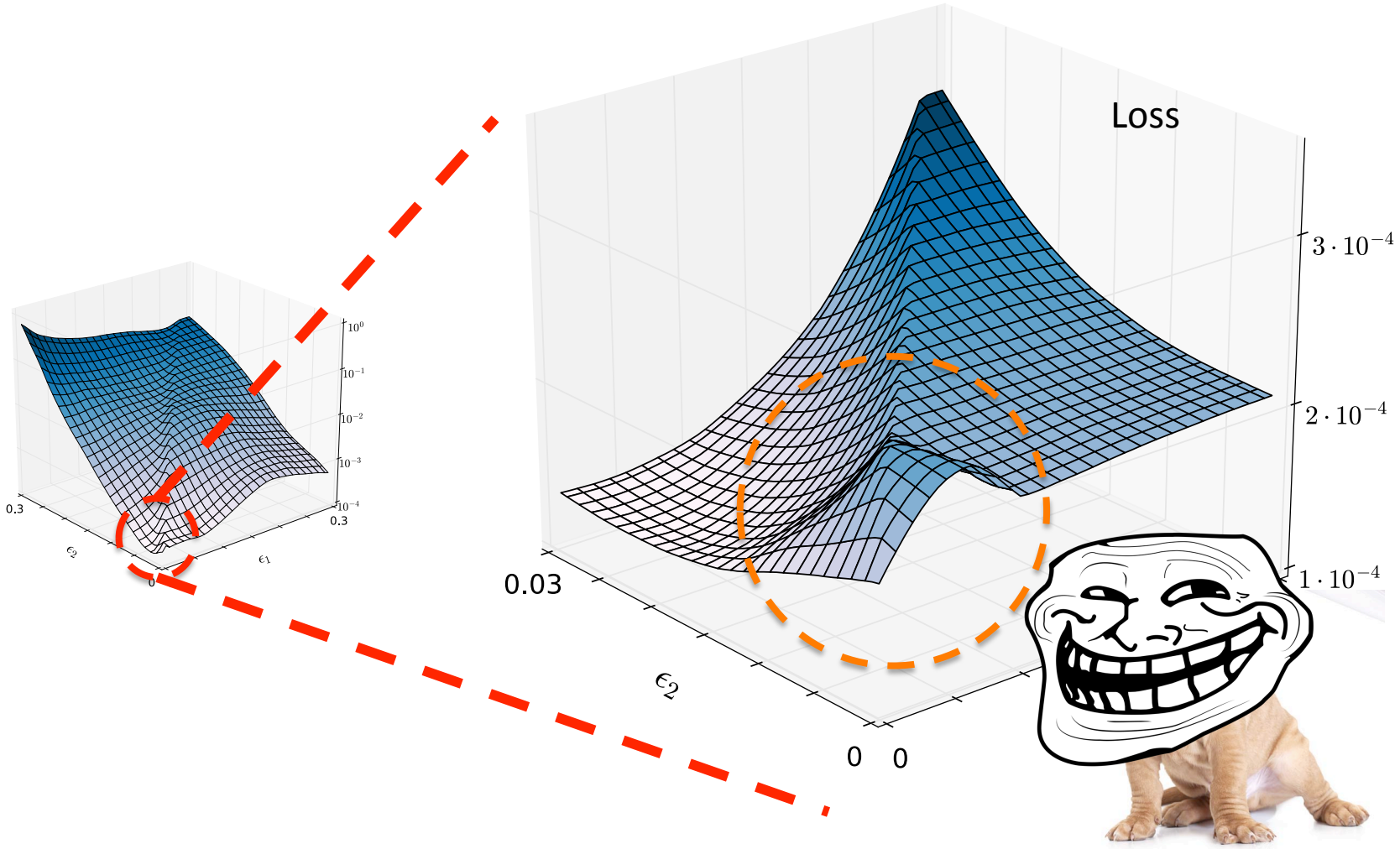
"Gradient Masking"



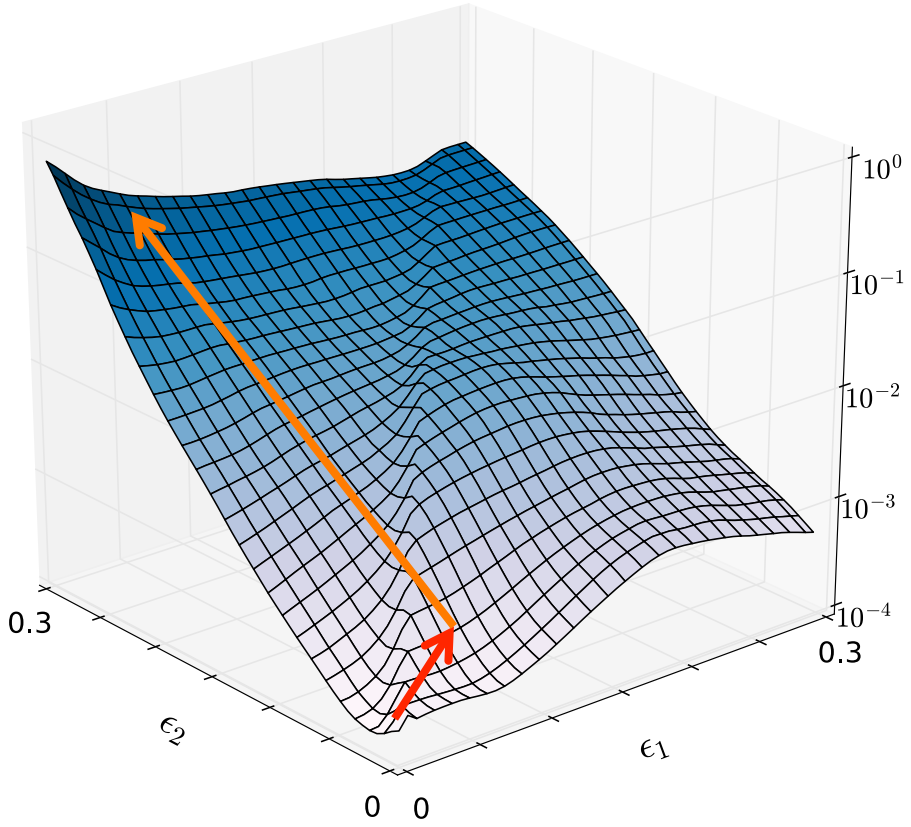
Loss of Adversarially Trained Model



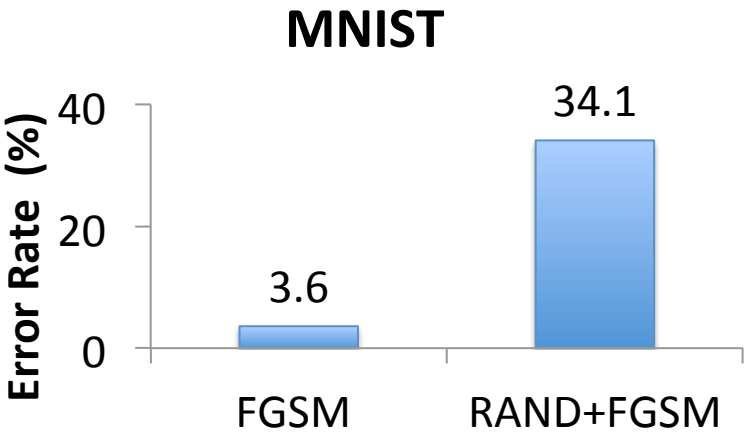
Loss of Adversarially Trained Model

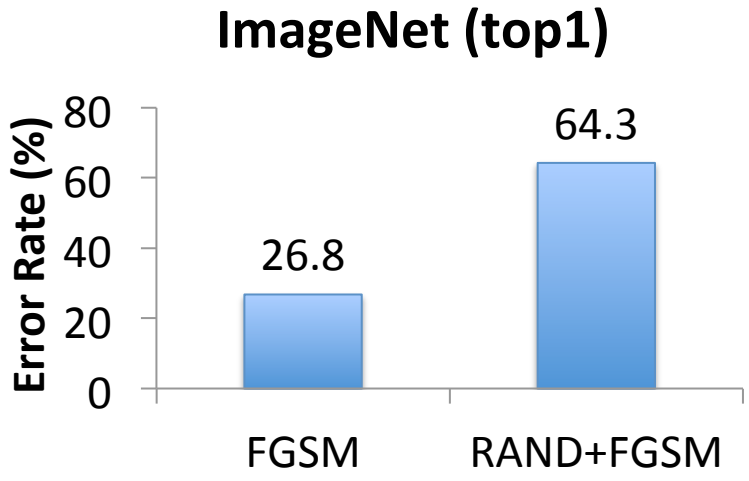


Simple One-Shot Attack: RAND+FGSM



1. Small random step
2. Step in direction of gradient

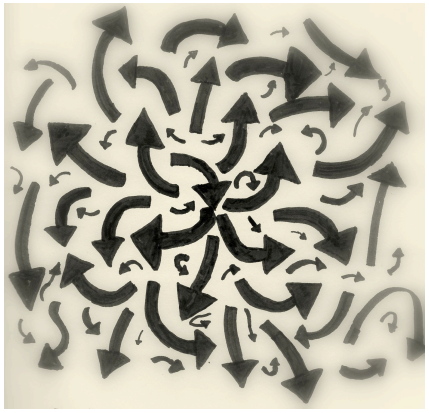




FGSM vs RAND+FGSM

- An improved one-shot attack even against non-defended models:
  - ≈ + **4%** error on MNIST
  - ≈ + **11%** error on ImageNet

- Adversarial training with RAND+FGSM
  - Doesn't work…
  - Are we stuck with adversarial training?



What's Wrong with Adversarial Training?

- Minimize

$$\text{loss}(x, y) + \underbrace{\text{loss}(x + \epsilon \cdot \texttt{sign}(\text{grad}), y)}$$

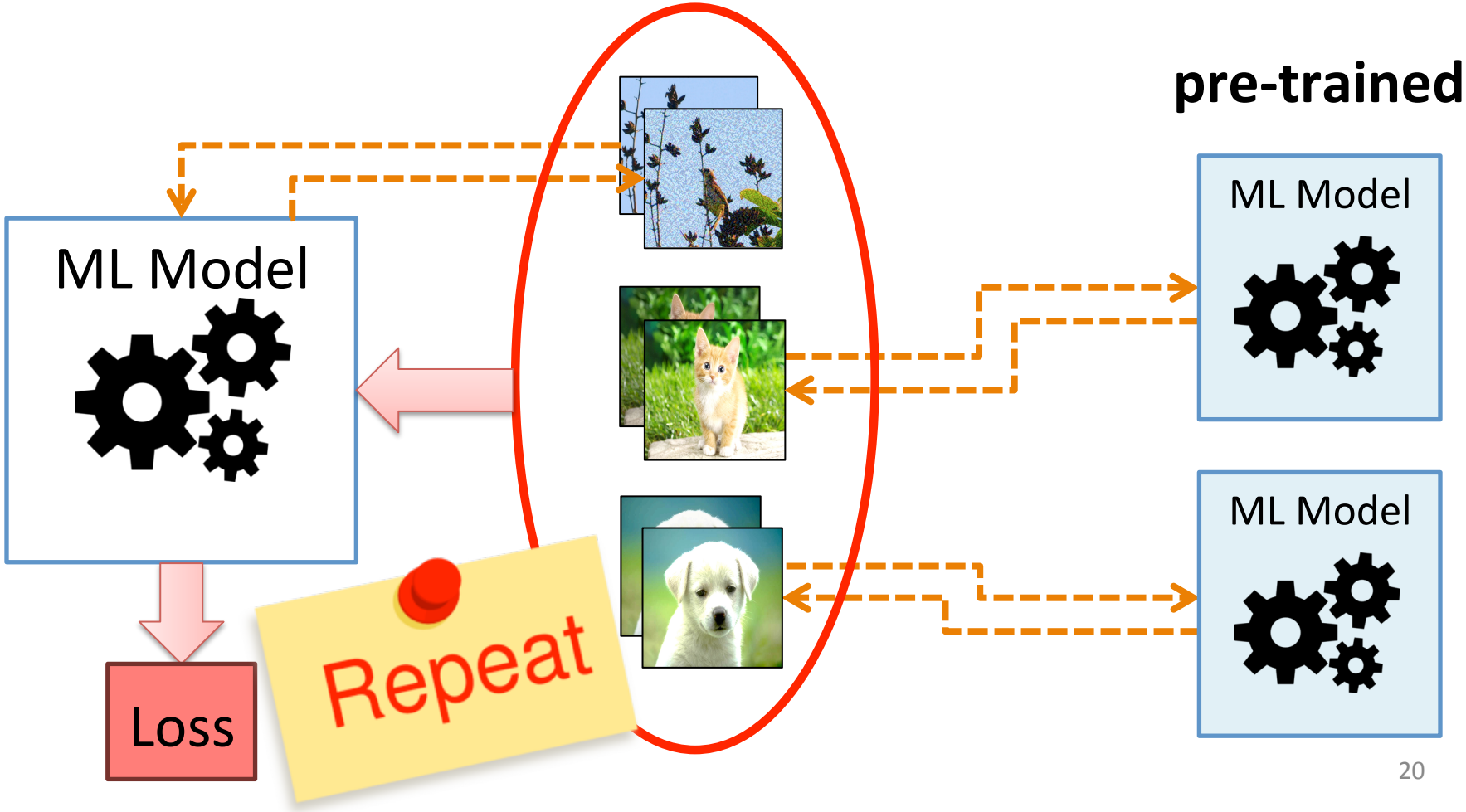
Small if:

1. The model is actually robust
2. *Or, the gradient points in a direction that is not adversarial*

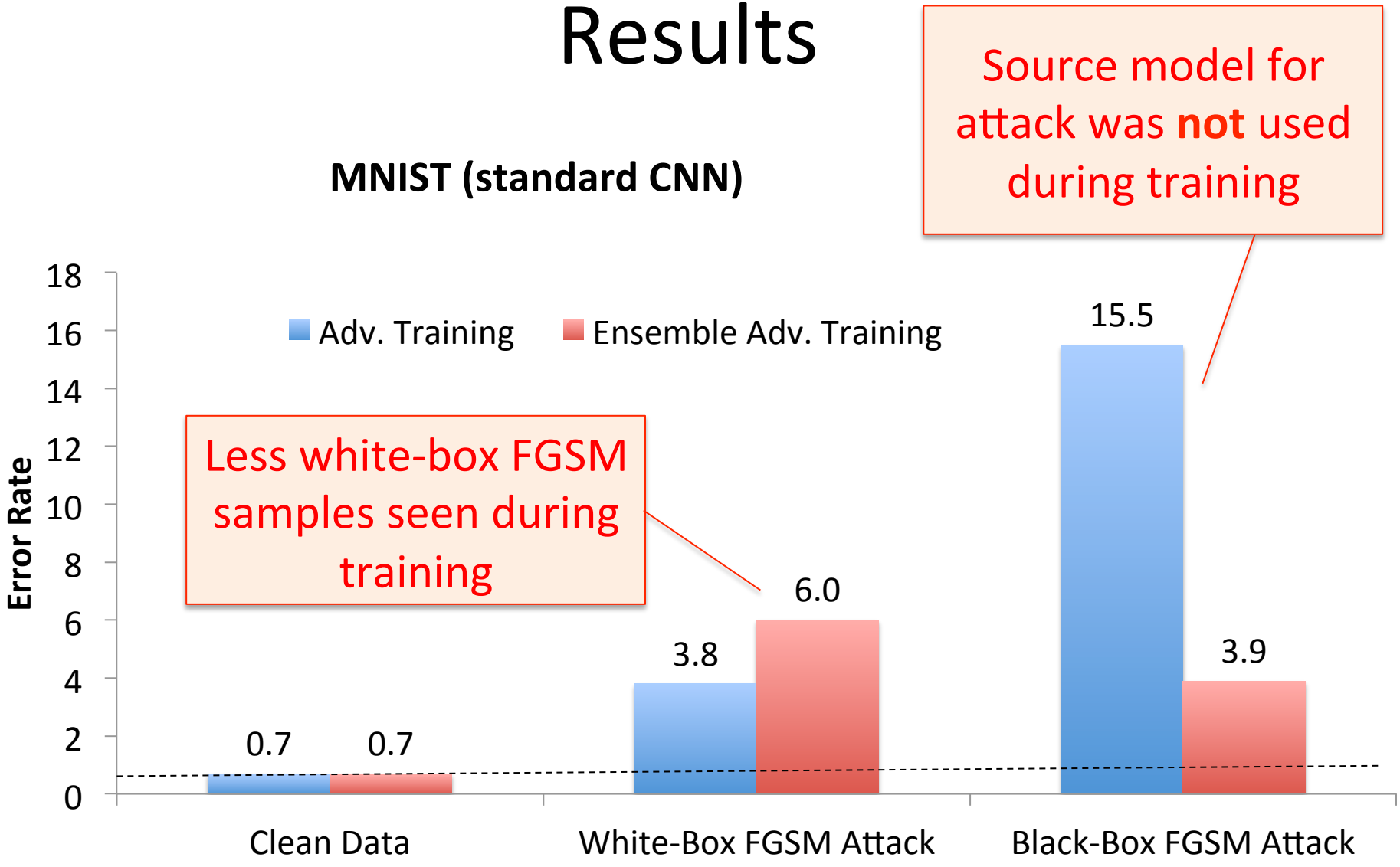
Degenerate Minimum

Ensemble Adversarial Training

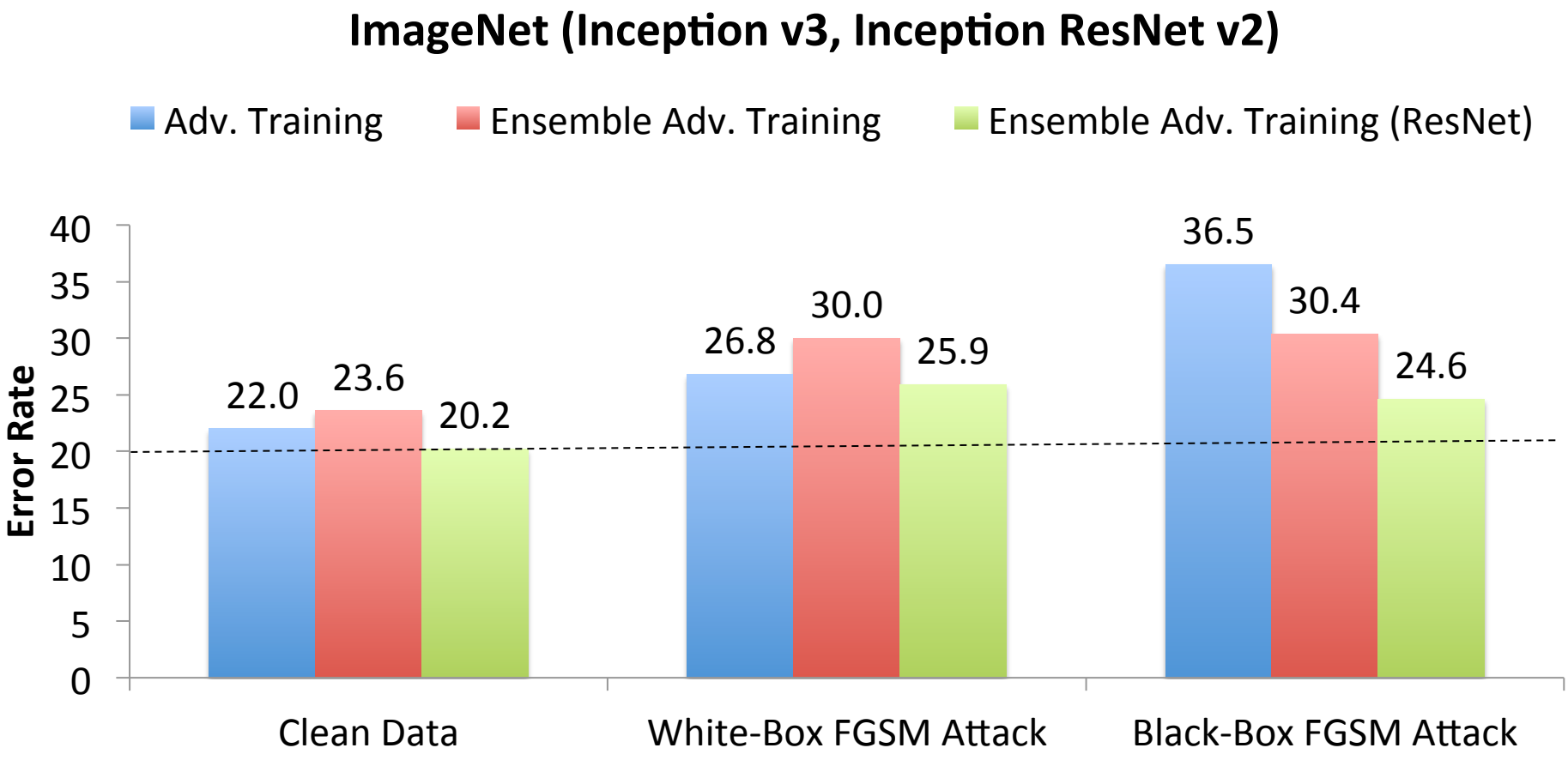
- How do we avoid these degenerate minima?



Results


MNIST (standard CNN)
Source model for attack was **not** used during training
Adv. Training
Ensemble Adv. Training
Error Rate
18
16
14
12
10
8
6
4
2
0
Less white-box FGSM samples seen during training
0.7
0.7
3.8
6.0
15.5
3.9
Clean Data
White-Box FGSM Attack
Black-Box FGSM Attack

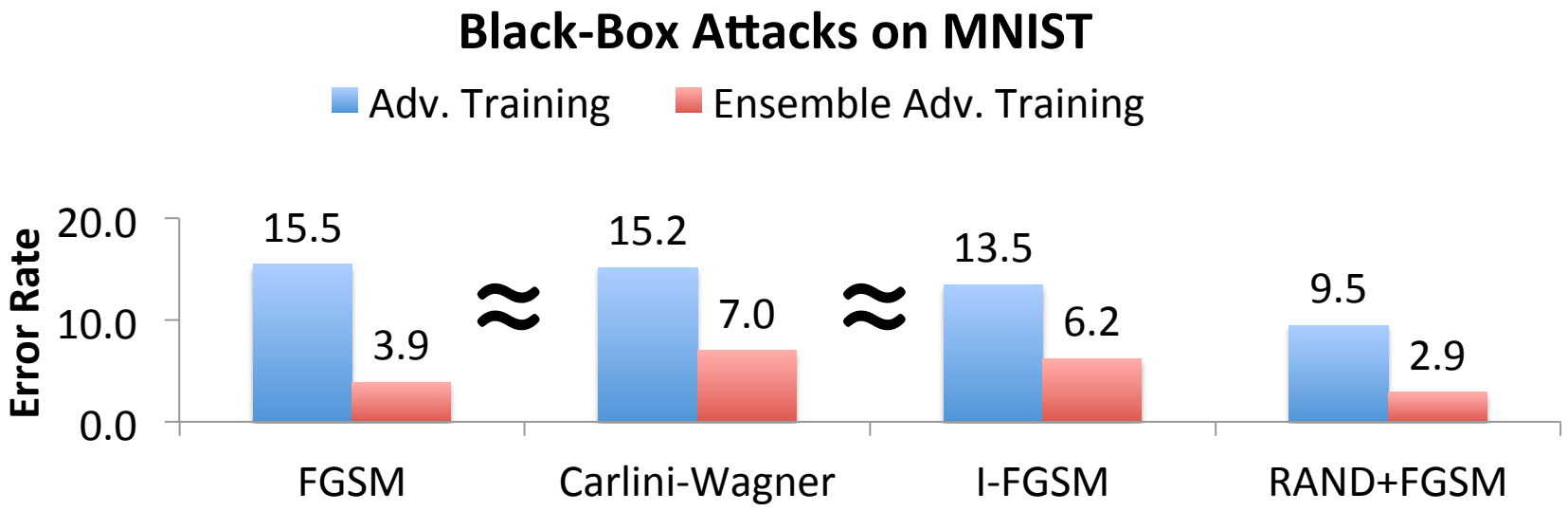
Results


ImageNet (Inception v3, Inception ResNet v2)
Adv. Training
Ensemble Adv. Training
Ensemble Adv. Training (ResNet)
Error Rate
40
35
30
25
20
15
10
5
0
22.0
23.6
20.2
26.8
30.0
25.9
36.5
30.4
24.6
Clean Data
White-Box FGSM Attack
Black-Box FGSM Attack

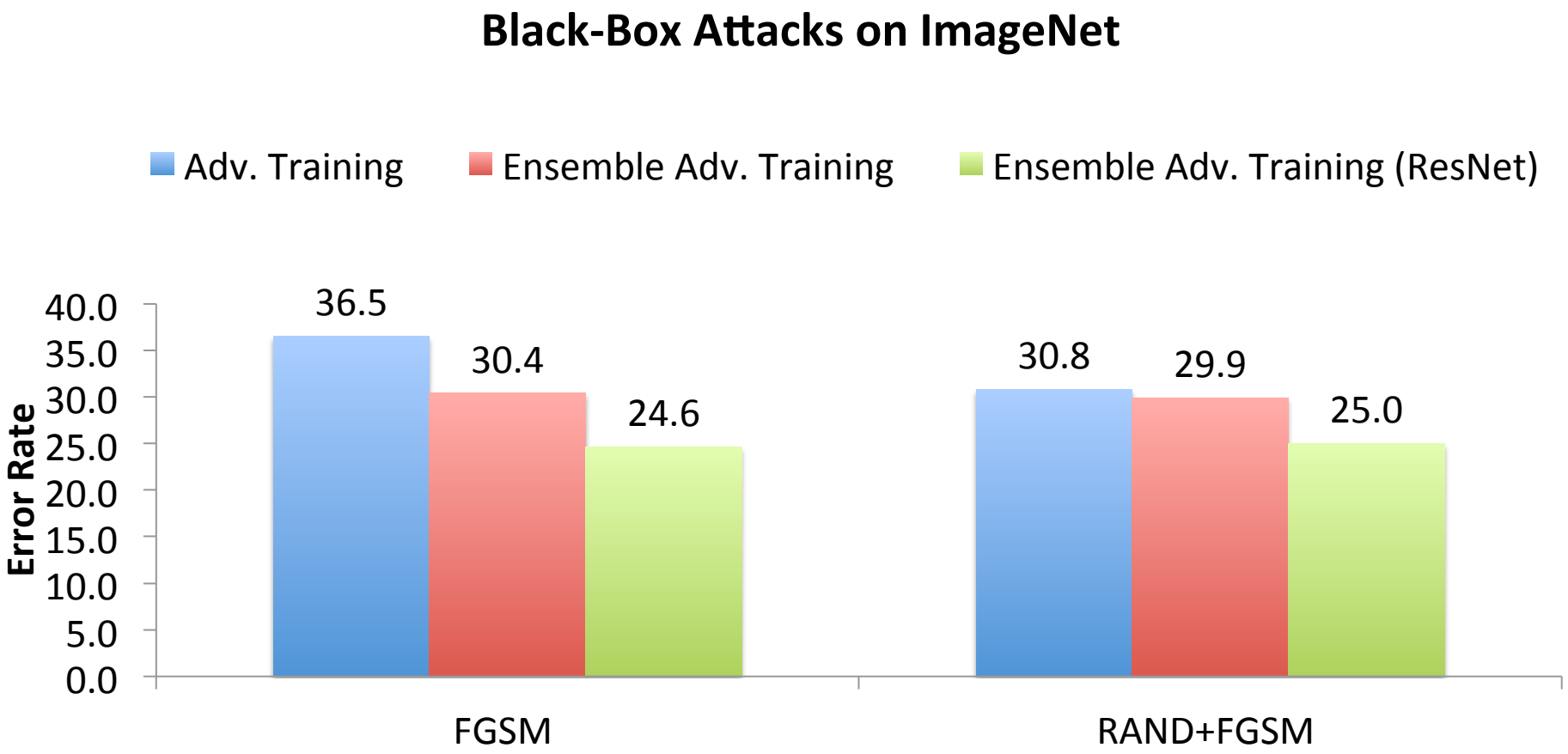
What about stronger attacks?

- Little to no improvement on **white-box** iterative and RAND+FGSM attacks!

- But, **improvements in black-box setting**!



What about stronger attacks?



Practical Considerations for Ensemble Adversarial Training

- **Pre-compute gradients** for pre-trained models
  - Lower per-batch cost than with adversarial training!

- **Randomize source model** in each batch
  - If `num_models % num_batches = 0`, we see the same adversarial examples in each epoch if we just rotate

- **Convergence is slower**

Standard Inception v3:	~150 epochs
Adversarial training:	~190 epochs
Ensemble adversarial training:	**~280 epochs**

Maybe because the task is actually hard?...

Takeaways

- Test defenses on black-box attacks
  - Distillation (Papernot et al. 2016, attack by Carlini et al. 2016)
  - Biologically Inspired Networks
    (Nayebi & Ganguli 2017, attack by Brendel & Bethge 2017)
  - Adversarial Training, and probably many others…
-
  « If you don't know where to go, just move at random. »
  — *Morgan Freeman* — *(or Dan Boneh)*
- Ensemble Adversarial Training can improve robustness to black-box attacks

Open Problems

- Better black-box attacks?
  - Using **ensemble** of source models? (Lin et al. 2017)
  - How much does ***oracle access*** to the model help?

- More efficient ensemble adversarial training?

- Can we say **anything** formal (and useful) about adversarial examples?

THANK YOU